Manuscript revised for *Science* August 18th, 2003

Hexapods Resurrected

(Technical comment on: "Hexapod Origins: Monophyletic or Paraphyletic?")

Frédéric Delsuc, Matthew J. Phillips and David Penny

The Allan Wilson Centre for Molecular Ecology and Evolution
Massey University, Palmerston North, New Zealand

Correspondence:

David Penny
The Allan Wilson Center for Molecular Ecology and Evolution
Massey University, Palmerston North, P O Box 11-222
New Zealand
Tel: +64 6 350 5033 / Fax: +64 6 350 5626
Courier Address: Institute of Molecular BioSciences, Science Tower D
E-mail: D.Penny@massey.ac.nz

Abstract

Nardi *et al.* (Science, 21 March 2003, 1887) suggested that extant hexapods might be diphyletic based on the analysis of amino acids sequences of mitochondrial genes. However, improved phylogenetic analyses of nucleotide sequences (RY-coding) from the same genes consistently retrieve the classical hypothesis of hexapod monophyly.

Text (741 words)

A recent paper (1) suggested, rather cautiously, that hexapods (insects plus collembolans in their dataset) might be diphyletic, rather than forming a monophyletic group. In their results, collembolans appeared well separated from other insects and emerged before crustaceans. Such an unexpected result has huge consequences for the interpretation of both morphological and developmental evolution in arthropods (2) and therefore deserves further scrutiny especially from the methodological point of view.

The authors drew their conclusions from maximum likelihood and Bayesian analyses at the amino acid level of four of the 12 mitochondrial proteins for both the original data set of 35 taxa and a reduced data set of 15 taxa. However, phylogenetic analyses of amino acids currently suffer from potential caveats. First, the currently available models of mitochondrial amino acid substitution are based on empirically deduced matrices from mammalian dominated sequence databases. Second, the maximum likelihood implementation used by the authors does not model the variation of rate across sites, known to be one of the most important parameter of the likelihood model (3). And third, bias in

nucleotide composition also affects the amino acid composition of the gene product causing potential problems for phylogenetic reconstruction (4).

Some of these pitfalls might be avoided by analyzing nucleotide sequences for which more realistic models of sequence evolution and powerful reconstruction methods are available. In particular, we have recently shown in the case of mammalian complete mitochondrial genomes (5) that it is possible to deal with saturation and base composition heterogeneity by recoding the nucleotides into purines (R) and pyrimidines (Y). This provided a solution to longstanding controversies concerning the position of the root of the mammalian tree (5).

Applying this strategy on nucleotides from the original Nardi et al. data set strongly suggests that by correcting for different artifacts it is possible to extract a useful historical signal. Unlike the authors (1), we were able to retrieve the bee (*Apis*) and the louse (*Heterodoxus*) within insects (Fig. 1). The artefactual position of these taxa as sister-groups of ticks was explained as being a consequence of their very high shared AT nucleotide composition reflected in their amino acid content (1). From our results, base composition heterogeneity seems to be more easily accommodated in phylogenetic reconstructions using nucleotides. More importantly, our analysis conforms with classical views of arthropod phylogeny: collembolans are sister-group of insects, and these monophyletic hexapods group with crustaceans into Pancrustacea (Fig. 1). One remaining problem with this tree concerns the paraphyly of myriapods induced by the nesting of the centipede (*Lithobius*) inside chelicerates.

However, as noted by the authors (1), their 35 taxa dataset also suffers from extreme rate heterogeneity among taxa, rendering it difficult to draw firm conclusions from. Thus, to check the collembolan position further, they reduced their data set to 15 taxa with more homogeneous rates and amino acid compositions. Despite their conservative analysis, they still reported collembolans outside both insects and crustaceans, rendering hexapods diphyletic. However, such a reduced dataset is particularly prone to systematic biases from low taxon sampling (6). In fact, deleting taxa with very anomalous rates and nucleotide composition can be helpful, but care is required not to delete taxa that then leave long isolated branches potentially leading to long branches attraction phenomena (7). More specifically, the inclusion of a single outgroup can have a strong impact on the phylogenetic reconstruction even in the absence of rate heterogeneity (8). In the case of placental mammal mitogenomics, taxon sampling has been shown to be a major source of phylogenetic error (9) and we found that increasing the number and diversity of taxa gave excellent agreement between nuclear and mitochondrial data (10).

To maximize taxon sampling, we constructed a well balanced 25 taxa data set designed to break isolated long branches (especially in the outgroup) without adding strong rate heterogeneity. Phylogenetic analyses of this nucleotide data set including RY-coded third codon positions produced a tree where Arthropoda, Pancrustacea, Hexapoda, Insecta and Pterygota all appear as monophyletic groups, though with variable support (Fig. 2). Moreover, this topology is much more compatible with the current views on arthropod phylogeny (11). The probability of randomly selecting a topology compatible with this prior hypothesis is so small (10) that it provides strong evidence in favor of its veracity.

Obviously, additional complete mitochondrial genomes are needed to strengthen the tree further, but with the data and methods currently at hand, the hypothesis of a common ancestry for extant hexapods cannot be rejected.

References and notes

1. F. Nardi *et al., Science* **299**, 1887 (2003).

2. R.H Thomas, *Science* **299**, 1854 (2003).

3. J. Sullivan, D.L. Swofford, *Syst. Biol.* **50**, 723 (2001).

4. P.G. Foster, D.A. Hickey, *J. Mol. Evol.* **48**, 284 (1999).

5. M.J. Phillips, D. Penny, *Mol. Phylogenet. Evol*. **28**, 171 (2003).

6. D.J. Zwickl, D.M. Hillis, *Syst. Biol.* **51**, 588 (2002).

7. M.D. Hendy, D. Penny, *Syst. Zool.* **38**, 297 (1989).

8. B.R. Holland, D. Penny, M.D. Hendy, *Syst. Biol*. **52**, 229 (2003).

9. H. Philippe, *J. Mol. Evol.* **45**, 712 (1997).

10. Y.-H. Lin *et al.*, *Mol. Biol. Evol*. **19**, 2060 (2002).

11. G. Giribet, G.D. Hedgecombe, W.C. Wheeler, *Nature* **413**, 157 (2001).

12. F. Ronquist, J.P. Huelsenbeck, *Bioinformatics* **19**, 1572 (2003).

13. D.L. Swofford, *PAUP* Sinauer Associates, Sunderland Massachusetts* (2002).

14. D. Posada, K.A. Crandall, *Bioinformatics* **14**, 817 (1998).

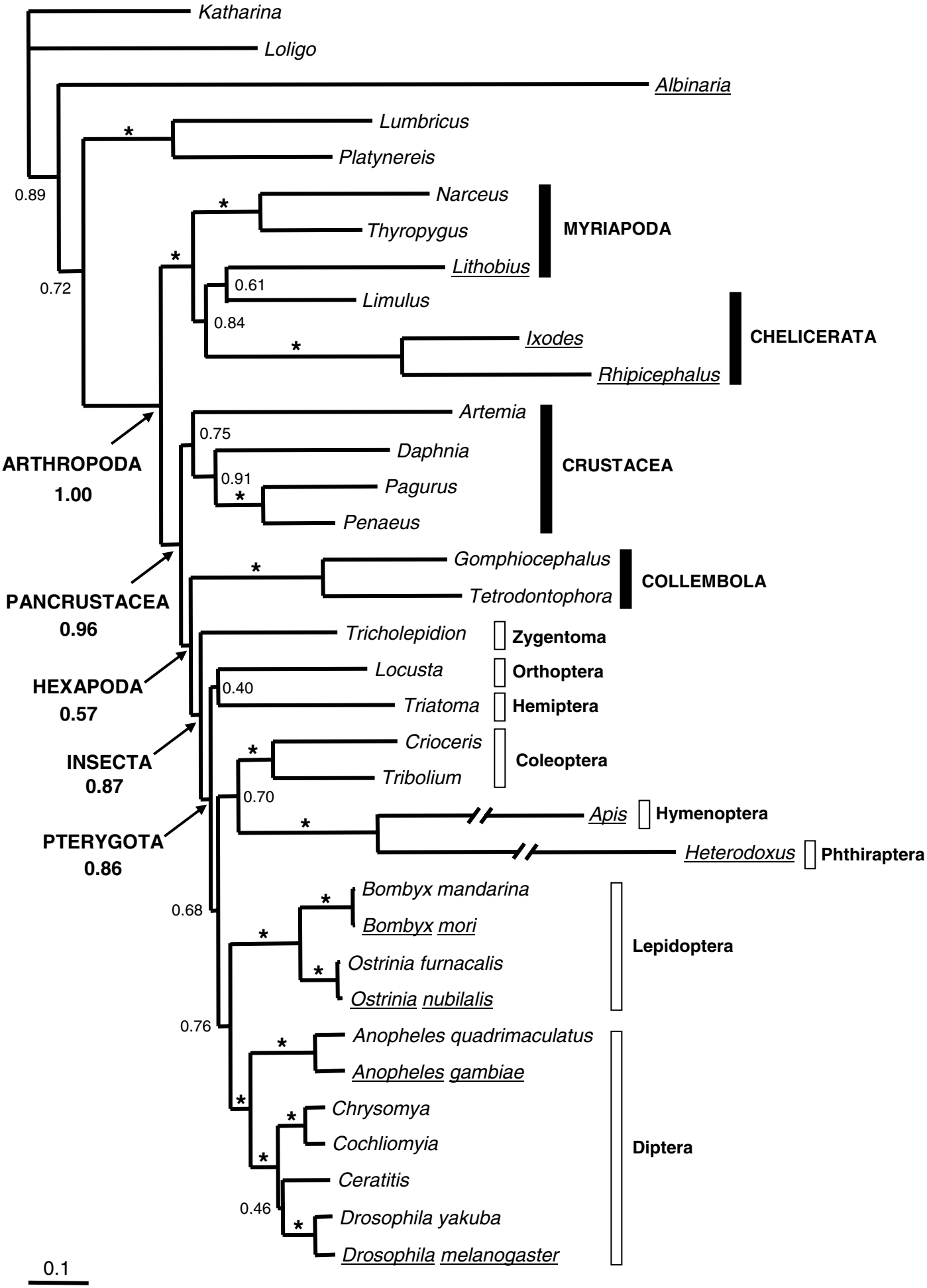
15. Francesco Nardi and colleagues kindly sent us their amino acid data set. Emmanuel Douzery provided helpful comments. Our data sets are available at http://awcmee.massey.ac.nz/downloads.htm. This work was supported by a Lavoisier

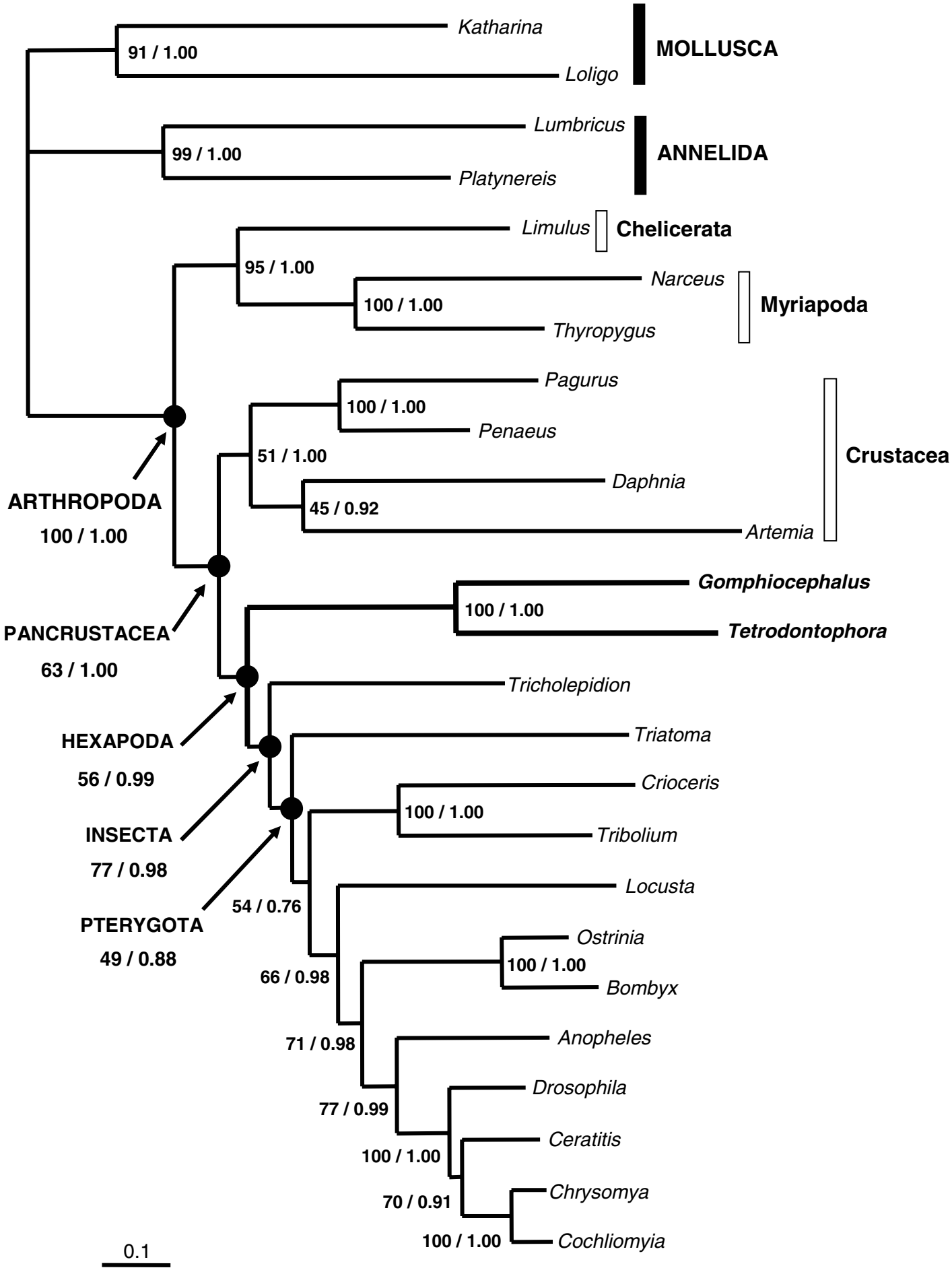
Postdoctoral Grant from the French Ministry of Foreign Affairs to FD and by the New Zealand Marsden Fund.

Figure legends

Fig. 1: Bayesian 50% majority rule consensus tree with associated branch lengths obtained using nucleotide sequences of COX1, COX2, COX3 and CYTB (3750 sites) corresponding to the 35 taxa data set of Nardi *et al.* (1). The first and third codon positions were RY-coded whereas second codon positions were kept as nucleotides. MrBayes version 3.0b4 (12) was used to perform a partitioned likelihood Bayesian search where three independent substitution models were attributed to each codon position: a two-state substitution model + I + Γ for RY coded first and third codon positions and a GTR + I + Γ model for second codon position nucleotides. Four incrementally heated MCMCMC were run for 500,000 generations sampling trees and parameters every 10 generations. The consensus tree was obtained from the 35,000 trees sampled after the initial burn in period. Values at nodes indicate Bayesian posterior probabilities (* = 1.00). Note that the terminal branch lengths leading to the bee (*Apis*) and the louse (*Heterodoxus*) have been reduced by a factor three. Underlined taxa are not included in the 25 taxa data set.

Fig. 2: Maximum likelihood (ML) phylogram obtained using nucleotide sequences of COX1, COX2, COX3 and CYTB for a 25 taxa data set (3777 sites). The third codon positions were RY-coded whereas first and second codon positions were kept as nucleotides. PAUP* version 4.0b10 (13) was used to perform a ML heuristic search under the best fitting GTR + I + Γ model and associated ML estimates of parameters as determined by Modeltest version 3.06 (14). A partitioned likelihood Bayesian search was carried out with MrBayes (12) using a GTR + I + Γ model for first and second codon position nucleotides and a two-state substitution model + I + Γ for the RY-coded third codon positions with the same parameter settings than in Fig. 1. Values at nodes indicate ML bootstrap proportions (100 replications) / Bayesian posterior probabilities. The two collembolans are figured in bold.


Katharina
Loligo
Albinaria
Lumbricus
Platynereis
Narceus
Thyropygus
MYRIAPODA
Lithobius
Limulus
CHELICERATA
Ixodes
Rhipicephalus
Artemia
Daphnia
Pagurus
Penaeus
CRUSTACEA
Gomphiocephalus
Tetrodontophora
COLLEMBOLA
Tricholepidion
Zygentoma
Locusta
Orthoptera
Triatoma
Hemiptera
Crioceris
Tribolium
Coleoptera
Apis
Hymenoptera
Heterodoxus
Phthiraptera
Bombyx mandarina
Bombyx mori
Ostrinia furnacalis
Ostrinia nubilalis
Lepidoptera
Anopheles quadrimaculatus
Anopheles gambiae
Chrysomya
Cochliomyia
Ceratitis
Drosophila yakuba
Drosophila melanogaster
Diptera
ARTHROPODA
1.00
PANCRUSTACEA
0.96
HEXAPODA
0.57
INSECTA
0.87
PTERYGOTA
0.86
0.89
0.72
0.61
0.84
0.75
0.91
0.40
0.70
0.68
0.76
0.46
0.1


Katharina
91 / 1.00
Loligo
MOLLUSCA
Lumbricus
99 / 1.00
Platynereis
ANNELIDA
Limulus
Chelicerata
95 / 1.00
Narceus
100 / 1.00
Thyropygus
Myriapoda
Pagurus
100 / 1.00
Penaeus
51 / 1.00
Daphnia
45 / 0.92
Artemia
Crustacea
ARTHROPODA
100 / 1.00
Gomphiocephalus
100 / 1.00
Tetrodontophora
PANCRUSTACEA
63 / 1.00
Tricholepidion
HEXAPODA
56 / 0.99
Triatoma
Crioceris
100 / 1.00
INSECTA
77 / 0.98
Tribolium
Locusta
54 / 0.76
PTERYGOTA
49 / 0.88
Ostrinia
100 / 1.00
66 / 0.98
Bombyx
Anopheles
71 / 0.98
Drosophila
77 / 0.99
Ceratitis
100 / 1.00
Chrysomya
70 / 0.91
100 / 1.00
Cochliomyia
0.1